Probabilistic Tracking and Reconstruction of 3D Human Motion in Monocular Video Sequences

Presentation of the thesis work of:
Hedvig Sidenbladh, KTH

Thesis opponent: Prof. Bill Freeman, MIT

Thesis supervisors
- Prof. Jan-Olof Eklundh, KTH
- Prof. Michael Black, Brown University

Collaborators
- Dr. David Fleet, Xerox PARC
- Prof. Dirk Ormoneit, Stanford University

A vision of the future from the past.

New York Worlds Fair, 1939
(Westinghouse Historical Collection)

Applications of computers looking at people
- Human-machine interaction
  - Robots
  - Intelligent rooms
- Video search
- Entertainment: motion capture for games, animation, and film.
- Surveillance

Technical Goal

Tracking a human in 3D

Why is it Hard?

The appearance of people can vary dramatically.
Why is it hard?

People can appear in arbitrary poses.

Structure is unobservable— inference from visible parts.

One solution:

- Use markers
- Use multiple cameras

Bregler and Malik ’98

- Brightness constancy cue
  - Insensitive to appearance
- Full-body required multiple cameras
- Single hypothesis

2D vs. 3D tracking

- Artist’s models...

Cham and Rehg ’99

- Single camera, multiple hypotheses
- 2D templates (no drift but view dependent)
  \[ l(x, t) = l(x + u, 0) + \eta \]
1999 state of art

Deutscher, North, Bascle, & Blake ‘00

- Multiple hypotheses
- Multiple cameras
- Simplified clothing, lighting and background

State of the Art.

Note: we can fake it with clever system design

Game videos...

Performance specifications

- No special clothing
- Monocular, grayscale, sequences (archival data)
- Unknown, cluttered, environment

Task: Infer 3D human motion from 2D image

Bayesian formulation

\[ p(\text{model} \mid \text{cues}) = \frac{p(\text{cues} \mid \text{model}) \cdot p(\text{model})}{p(\text{cues})} \]

1. Need a constraining likelihood model that is also invariant to variations in human appearance.
2. Need a prior model of how people move.
3. Posterior probability: Need an effective way to explore the model space (very high dimensional) and represent ambiguities.

System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

Simple Body Model

* Limbs are truncated cones
* Parameter vector of joint angles and angular velocities = \( \phi \)

Multiple Hypotheses

- Posterior distribution over model parameters often multi-modal (due to ambiguities)
- Represent whole distribution:
  - sampled representation
  - each sample is a pose
  - predict over time using a particle filtering approach

Particle Filter

\[ p(\phi_{t-1} \mid I_{t-1}) \rightarrow \text{Temporal dynamics} \rightarrow p(\phi_{t} \mid \phi_{t-1}) \]
\[ p(\phi_{t} \mid \phi_{t-1}) \rightarrow \text{Sample} \rightarrow \text{normalize} \rightarrow p(I_{t} \mid \phi_{t}) \]

Temporal dynamics: \( p(\phi_{t} \mid \phi_{t-1}) \)
Sample: \( \phi_{t} \)

Problem: Expensive representation of posterior!
Approaches to solve problem:
- Lower the number of samples. (Deutsher et al., CVPR00)
- Represent the space in other ways (Choo and Fleet, ICCV01)
System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

Key Idea #1 (Likelihood)

1. Use the 3D model to predict the location of limb boundaries (not necessarily features) in the scene.
2. Compute various filter responses steered to the predicted orientation of the limb.
3. Compute likelihood of filter responses using a statistical model learned from examples.

Edge Detection?

- Probabilistic model?
- Under/over-segmentation, thresholds, …

Edge Filters

Normalized derivatives of Gaussians (Lindeberg, Granlund and Knutsson, Porona, Freeman & Adelson, …)

Edge filter response steered to limb orientation:

\[ f^* (x, \theta, \sigma) = \sin \theta f_x(x, \sigma) + \cos \theta f_y(x, \sigma) \]

Example Training Images

- Filter responses steered to arm orientation.
**Edge Distributions**

Edge response steered to model edge:

$$f_c(x, \theta, \sigma) = \sin \theta f_x(x, \sigma) + \cos \theta f_y(x, \sigma)$$

Similar to Konishi et al., CVPR 99

---

**Edge Likelihood Ratio**

Edge response

Likelihood ratio

---

**Other Cues**

Ridges

Motion

I(x, t)

I(x+u, t+1)

---

**Ridge Distributions**

Ridge response steered to limb orientation

$$f_r(x, \theta, \sigma) = |\sin \theta f_x(x, \sigma) + \cos \theta f_y(x, \sigma)| - |\cos \theta f_x(x, \sigma) + \sin \theta f_y(x, \sigma)|$$

Ridge response only on certain image scales!

---

**Motion distributions**

Different underlying motion models

---

**Likelihood Formulation**

- Independence assumptions:
  - Cues: $$p(\text{image} | \text{model}) = p(\text{cue1} | \text{model}) p(\text{cue2} | \text{model})$$
  - Spatial: $$p(\text{image} | \text{model}) = \Pi_{x \in \text{image}} p(\text{image}(x) | \text{model})$$
  - Scales: $$p(\text{image} | \text{model}) = \Pi_{\sigma \in \text{scale}} p(\text{image}(\sigma) | \text{model})$$

- Combines cues and scales!
- Simplification, in reality there are dependencies
The power of cue combination

Using edge cues alone

Using ridge cues alone

Using flow cue alone

Using edge, ridge, and motion cues together

Key Idea #2

\[ p(\text{image} \mid \text{foreground, background}) \propto p(\text{foreground part of image} \mid \text{foreground}) p(\text{foreground part of image} \mid \text{background}) \]
**Likelihood**

\[ p(\text{image} | \text{fore, back}) = \prod_{\text{fore pixels}} p(\text{image} | \text{fore}) \prod_{\text{back pixels}} p(\text{image} | \text{back}) \]

\[ = \prod_{\text{fore pixels}} p(\text{image} | \text{back}) \prod_{\text{fore pixels}} p(\text{image} | \text{fore}) \]

\[ = \text{const} \prod_{\text{fore pixels}} p(\text{image} | \text{fore}) \prod_{\text{fore pixels}} p(\text{image} | \text{back}) \]

- Foreground pixels
- Background pixels

**System components**

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

**The Prior term**

Bayesian formulation:

\[ p(\text{model} | \text{cue}) \propto p(\text{cue} | \text{model}) p(\text{model}) \]

- Need a constraining likelihood model that is also invariant to variations in human appearance
- Need a good model of how people move

**Very general model**

- Constant velocity motions
- Not constrained by how people tend to move.

**Constant velocity model**

- All DOF in the model parameter space, \( \phi \), independent
- Angles are assumed to change with constant speed
- Speed and position changes are randomly sampled from normal distribution

**Tracking an Arm**

Moving camera, constant velocity model

1500 samples
- 2 min/frame
Self Occlusion

Constant velocity model

1500 samples
~2 min/frame

System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

Very specific model

- Only handles people walking.
- Very powerful constraint on human motion.

Models of Human Dynamics

- Action-specific model - Walking
  - Training data: 3D motion capture data
  - From training set, learn mean cycle and common modes of deviation (PCA)

Walking Person

Mean cycle Small noise Large noise

No likelihood

* how strong is the walking prior? (or is our likelihood doing anything?)
System components

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Very general model
  - Very specific model
  - Example-based model

Example-based model

- Take lots of training data.
- Use “snippets” of the data as models for how people are likely to move.

Example-based model

Ten samples from the prior, drawn using approximate probabilistic tree search.

Tracking with only 300 particles.

Lessons Learned

- Representation for probabilistic analysis.
  - Probabilistic (Bayesian) framework allows
  - Integration of information in a principled way
  - Modeling of priors
  - Particle filtering allows
    - Multi-modal distributions
    - Tracking with ambiguities and non-linear models
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
Lessons Learned

- Representation for probabilistic analysis.
- Models for human appearance (likelihood term).
- Models for human motion (prior term).
  - Explored 3 different models; analyzed the tradeoffs between each.

Bayesian Inference

Exploit cues in the images. Learn likelihood models:

\[ p(\text{image cue} \mid \text{model}) \]

Build models of human form and motion. Learn priors over model parameters:

\[ p(\text{model}) \]

Represent the posterior distribution:

\[ p(\text{model} \mid \text{cue}) \propto p(\text{cue} \mid \text{model}) p(\text{model}) \]